CHAPTER 7 - PROPERTIES of EXPECTATION

7.1 - Introduction

In this chapter, we develop and exploit additional properties of expected values

Reminder:

The expected value of a random variable X is defined by:

$$E[X] = \sum_x x p(x)$$

where X is a discrete random variable with probability mass function (pdf) $p(x)$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx.$$

$-\infty$

Where X is a continuous random variable with probability density function $f(x)$.

7.2 - Expectations of Sums of Random Variables

Example: 2a

An accident occurs at a point X that is uniformly distributed on a road of length L.
At the time of the accident, an ambulance is at a location Y, that is also uniformly distributed on the road.
- Assume that X & Y are independent

find the expected distance between the ambulance and the point of the accident

→ Solution :

Need to find $E[|X - Y|]$, since the joint density function of X and Y is :

$$f(x,y) = \frac{1}{L^2} \quad 0 < x < L, \; 0 < y < L$$

Follows from Proposition 2.1

Proposition 2.1. If X and Y have a joint probability mass function $p(x,y)$, then

$$E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$$

If X and Y have a joint probability density function $f(x,y)$, then

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy$$

$$E[|X - Y|] = \frac{1}{L^2}\int_0^L \int_0^L |x - y|\, dy\, dx$$

Now,

$$\begin{aligned}\int_0^L |x - y| dy &= \int_0^x (x - y) dy + \int_x^L (y - x) dy \\ &= \frac{x^2}{2} + \frac{L^2}{2} - \frac{x^2}{2} - x(L - x) \\ &= \frac{L^2}{2} + x^2 - xL\end{aligned}$$

Therefore,

$$\begin{aligned}E[|X - Y|] &= \frac{1}{L^2}\int_0^L \left(\frac{L^2}{2} + x^2 - xL\right) dx \\ &= \frac{L}{3}\end{aligned}$$

■

For an important application of Proposition 2.1, suppose that $E[X]$ and $E[Y]$ are both finite and let $g(X, Y) = X + Y$. Then, in the continuous case,

$$\begin{aligned}E[X + Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x + y) f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x, y)\, dy\, dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy \\ &= E[X] + E[Y]\end{aligned}$$

The same result holds in general; thus, whenever $E[X]$ and $E[Y]$ are finite,

$$E[X + Y] = E[X] + E[Y] \tag{2.1}$$

Using Equation (2.1), we may show by a simple induction proof that if $E[X_i]$ is finite for all $i = 1, \ldots, n$, then

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] \tag{2.2}$$

Equation (2.2) is an extremely useful formula whose utility will now be illustrated by a series of examples.

> The sample mean

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, having distribution function F & expected value μ

Such a sequence of random

variables" is said to constitute a sample from the distribution F.

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the *sample mean*. Compute $E[\overline{X}]$.

Solution.

$$\begin{aligned} E[\overline{X}] &= E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[X_i] \\ &= \mu \quad \text{since } E[X_i] \equiv \mu \end{aligned}$$

That is, the expected value of the sample mean is μ, the mean of the distribution. When the distribution mean μ is unknown, the sample mean is often used in statistics to estimate it. ■

- EXAMPLE 2d - Boole's inequality

- Let $A_1, \ldots, A_n$ denote events, and define the indicator variables X_i, $i = 1, \ldots, n$ by

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Let

$$X = \sum_{i=1}^{n} X_i$$

so X denotes the number of the events A_i, that occur.

Finally let,

$$Y = \begin{cases} 1 & X \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

so that Y is equal to 1 if at least one of the A_i occurs and is 0 otherwise.

$\therefore$ $X \geq Y$ so $E[X] \geq E[Y]$

But since,

$$E[X] = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} P(A_i)$$

Σ

$$E[Y] = P(\text{at least one of the } A_i \text{ occurs})$$
$$= P\left(\bigcup_{i=1}^{n} A_i\right)$$

we obtain Boole's inequality.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

The next 3 examples show how (Equation 2.2) can be used to calculate the expected values of binomial, negative binomial & hypergeometric random variable

↳ see pg 316 - 317.

Note: See book for more details.

7.4 - Covariance, Variance of Sums, and Correlations

The following propositions show that the expectation of a product of independent random variables is equal to the product of their expectation

Proposition 4.1. If X and Y are independent, then, for any functions h and g,

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof. Suppose that X and Y are jointly continuous with joint density $f(x, y)$. Then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)\,dx \\ &= E[h(Y)]E[g(X)] \end{aligned}$$

The proof in the discrete case is similar. □

Just as the expected value and the variance of a single random variable give us information about that random variable, so does the covariance between two random variables give us information about the relationship between the random variables.

Definition:

The covariance between X & Y.

denoted by $Cov(X, Y)$ is defined by:

$$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

Upon expanding the right side of the preceding definition, we see that

$$\begin{aligned} \text{Cov}(X, Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Note that if X and Y are independent, then, by Proposition 4.1, $\text{Cov}(X, Y) = 0$. However, the converse is not true. A simple example of two dependent random variables X and Y having zero covariance is obtained by letting X be a random variable such that

$$P\{X = 0\} = P\{X = 1\} = P\{X = -1\} = \frac{1}{3}$$

and defining

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0 \end{cases}$$

Now, $XY = 0$, so $E[XY] = 0$. Also, $E[X] = 0$. Thus,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0$$

However, X and Y are clearly not independent.

The following proposition lists some of the properties of covariance.

Proposition 4.2.

(i) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

(ii) $\text{Cov}(X, X) = \text{Var(X)}$

(iii) $\text{Cov}(aX, Y) = a\,\text{Cov}(X, Y)$

(iv) $$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

Notes: Proofs can be found pg 338.

+ EXAMPLES

[EXAMPLE 4c pg 341]

7.5 - Conditional Expectation

Definition:

Recall that if X & Y are jointly discrete random variables, then the conditional probability mass function of X, given that $Y=y$, is defined, for all y such that $P(Y=y) > 0$, by:

$$p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$$

It is therefore natural to define, in this case, the conditional expectation of X given that $Y = y$, for all values of y such that $p_Y(y) > 0$, by

$$E[X|Y = y] = \sum_x xP\{X = x|Y = y\}$$
$$= \sum_x xp_{X|Y}(x|y)$$

Lo 7.5.2 - Computing Expectations by Conditioning

- Let us denote by $E[X|Y]$ that function of the random variable Y whose value at $Y=y$ is $E[X|Y=y]$
- Note that $E[X|Y]$ is itself a random variable.

Important property:

Proposition 5.1.

$$E[X] = E[E[X|Y]] \tag{5.1}$$

If Y is a discrete random variable, then Equation (5.1) states that

$$E[X] = \sum_y E[X|Y=y]P\{Y=y\} \tag{5.1a}$$

whereas if Y is continuous with density $f_Y(y)$, then Equation (5.1) states

$$E[X] = \int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)\,dy \tag{5.1b}$$

We now give a proof of Equation (5.1) in the case where X and Y are both discrete random variables.

Example 5C

A miner is trapped in a mine containing 3 doors

- 1st door leads to a tunnel that will take him to safety after 3 hours.
- 2nd door leads to a tunnel that will return him to the mine after 5 hours of travel.
- 3rd door leads to a tunnel that will return him to the mine after 7 hours.

- If we assume that the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until he reaches safety.

Solution:

- Let X denote the amount of time (hours) until miner reaches safety
- Let Y denote the door he initially chooses.

∴

$$
\begin{aligned}
E[X] &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} \\
&\quad + E[X|Y=3]P\{Y=3\} \\
&= \frac{1}{3}(E[X|Y=1] + E[X|Y=2] + E[X|Y=3])
\end{aligned}
$$

However,

$$
\begin{aligned}
E[X|Y=1] &= 3 \\
E[X|Y=2] &= 5 + E[X] \\
E[X|Y=3] &= 7 + E[X]
\end{aligned}
\tag{5.3}
$$

> To understant equation (5.3) let take for example

$E[X|Y=2]$

→ If the miner chooses the second door, he spends 5 hours in the tunnel and then returns to his cell.

BUT, once he returns to his cell the problem is as before; thus his expected additional time until safety is just $E[X]$.

∴

$$E[X \mid Y = 2] = 5 + E[X].$$

Note: The argument behind the other equalities in Equation (5.3) is similar.

$$E[X] = \frac{1}{3}(3 + 5 + E[X] + 7 + E[X])$$

or

$E[X] = 15$

Note: More examples can be found through the chapter.

7.5.3 - Computing Probabilities by Conditioning

Example 5k - The best-prize problem

Suppose that we are to be presented with n distinct prizes, in sequence. After being presented with a prize, we must immediately decide whether to accept it or to reject it and consider the next prize.

- The only information we are given when deciding whether to accept a prize is the relative rank of the prize compared to ones already seen

For instance, when the 5th prize is presented, we learn how it compares with the four prizes we've already seen.

- Suppose that once a prize is rejected, it is lost, and that our objective is to maximize the probability of obtaining the best prize. Assuming that all n! orderings of the prize are equally likely, how well can we do?

Solution. Rather surprisingly, we can do quite well. To see this, fix a value k, $0 \leq k < n$, and consider the strategy that rejects the first k prizes and then accepts the first one that is better than all of those first k. Let $P_k(\text{best})$ denote the probability that the best prize is selected when this strategy is employed. To compute this probability, condition on X, the position of the best prize. This gives

$$\begin{aligned} P_k(\text{best}) &= \sum_{i=1}^{n} P_k(\text{best}|X=i)P(X=i) \\ &= \frac{1}{n}\sum_{i=1}^{n} P_k(\text{best}|X=i) \end{aligned}$$

Now, on the one hand, if the overall best prize is among the first k, then no prize is ever selected under the strategy considered. That is,

$$P_k(\text{best}|X=i) = 0 \quad \text{if } i \leq k$$

On the other hand, if the best prize is in position i, where $i > k$, then the best prize will be selected if the best of the first $i-1$ prizes is among the first k (for then none of the prizes in positions $k+1, k+2, \ldots, i-1$ would be selected). But, conditional on the best prize being in position i, it is easy to verify that all possible orderings of the other prizes remain equally likely, which implies that each of the first $i-1$ prizes is equally likely to be the best of that batch. Hence, we have

$$\begin{aligned} P_k(\text{best}|X=i) &= P\{\text{best of first } i-1 \text{ is among the first } k|X=i\} \\ &= \frac{k}{i-1} \quad \text{if } i > k \end{aligned}$$

From the preceding, we obtain

$$\begin{aligned} P_k(\text{best}) &= \frac{k}{n}\sum_{i=k+1}^{n}\frac{1}{i-1} \\ &\approx \frac{k}{n}\int_{k+1}^{n}\frac{1}{x-1}\,dx \\ &= \frac{k}{n}\log\left(\frac{n-1}{k}\right) \\ &\approx \frac{k}{n}\log\left(\frac{n}{k}\right) \end{aligned}$$

Now, if we consider the function

$$g(x) = \frac{x}{n}\log\left(\frac{n}{x}\right)$$

then

$$g'(x) = \frac{1}{n}\log\left(\frac{n}{x}\right) - \frac{1}{n}$$

so

$$g'(x) = 0 \Rightarrow \log\left(\frac{n}{x}\right) = 1 \Rightarrow x = \frac{n}{e}$$

Thus, since $P_k(\text{best}) \approx g(k)$, we see that the best strategy of the type considered is to let the first n/e prizes go by and then accept the first one to appear that is better than all of those. In addition, since $g(n/e) = 1/e$, the probability that this strategy selects the best prize is approximately $1/e \approx .36788$.

Remark. Most people are quite surprised by the size of the probability of obtaining the best prize, thinking that this probability would be close to 0 when n is large. However, even without going through the calculations, a little thought reveals that the probability of obtaining the best prize can be made reasonably large. Consider the strategy of letting half of the prizes go by and then selecting the first one to appear that is better than all of those. The probability that a prize is actually selected is the probability that the overall best is among the second half, and this is $\frac{1}{2}$. In addition, given that a prize is selected, at the time of selection that prize would have been

the best of more than $n/2$ prizes to have appeared and would thus have probability of at least $\frac{1}{2}$ of being the overall best. Hence, the strategy of letting the first half of all prizes go by and then accepting the first one that is better than all of those prizes has a probability greater than $\frac{1}{4}$ of obtaining the best prize. ■

7.5.4 - Conditional Variance

Just as we have defined the conditional expectation of X given the value of Y, we can also define the conditional variance of X given that $Y = y$:

$$\text{Var}(X|Y) \equiv E[(X - E[X|Y])^2|Y]$$

That is, $\text{Var}(X|Y)$ is equal to the (conditional) expected square of the difference between X and its (conditional) mean when the value of Y is given. In other words, $\text{Var}(X|Y)$ is exactly analogous to the usual definition of variance, but now all expectations are conditional on the fact that Y is known.

There is a very useful relationship between $\text{Var}(X)$, the unconditional variance of X, and $\text{Var}(X|Y)$, the conditional variance of X given Y, that can often be applied to compute $\text{Var}(X)$. To obtain this relationship, note first that, by the same reasoning that yields $\text{Var}(X) = E[X^2] - (E[X])^2$, we have

$$\text{Var}(X|Y) = E[X^2|Y] - (E[X|Y])^2$$

so

$$\begin{aligned} E[\text{Var}(X|Y)] &= E[E[X^2|Y]] - E[(E[X|Y])^2] \\ &= E[X^2] - E[(E[X|Y])^2] \end{aligned} \tag{5.9}$$

Also, since $E[E[X|Y]] = E[X]$, we have

$$\text{Var}(E[X|Y]) = E[(E[X|Y])^2] - (E[X])^2 \tag{5.10}$$

Hence, by adding Equations (5.9) and (5.10), we arrive at the following proposition.

Proposition 5.2. The conditional variance formula

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

7.6 - Conditional Expectation & Prediction

EXAMPLE 6a

Suppose that the son of a man of height x (in inches) attains a height that is normally distributed with mean $x + 1$ and variance 4. What is the best prediction of the height at full growth of the son of a man who is 6 feet tall?

Solution. Formally, this model can be written as

$$Y = X + 1 + e$$

where e is a normal random variable, independent of X, having mean 0 and variance 4. The X and Y, of course, represent the heights of the man and his son, respectively. The best prediction $E[Y|X = 72]$ is thus equal to

$$\begin{aligned} E[Y|X = 72] &= E[X + 1 + e|X = 72] \\ &= 73 + E[e|X = 72] \\ &= 73 + E(e) \quad \text{by independence} \\ &= 73 \end{aligned}$$

■

7.7 - Moment Generating Functions

The moment generating function $M(t)$ of the random variable X is defined for all real values of t by

$$M(t) = E[e^{tX}]$$
$$= \begin{cases} \sum_x e^{tx} p(x) & \text{if } X \text{ is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx & \text{if } X \text{ is continuous with density } f(x) \end{cases}$$

We call $M(t)$, the moment generating function because all of the moments of X can be obtained by successively differentiating $M(t)$ and then evaluating the result at $t=0$.

eg

$$M'(t) = \frac{d}{} E[e^{tx}]$$

$$= E\left[\frac{d}{dt} e^{tx}\right]$$

$$= E\left[x e^{tx}\right]$$

In general, the n^{th} derivative of $M(t)$ is given by:

$$M^n(t) = E\left[X^n e^{tx}\right] \quad n \geq 1$$

implying that

$$M^n(0) = E\left[X^n\right] \quad n \geq 1$$

Note:

- EXAMPLE 7a - BINOMIAL pg 370
- EXAMPLE 7b - Poisson. pg 370

- EXAMPLE 7c - Exponential pg 371

- EXAMPLE 7d - Normal pg 371

TABLE 7.1: DISCRETE PROBABILITY DISTRIBUTION

	Probability mass function, $p(x)$	Moment generating function, $M(t)$	Mean	Variance
Binomial with parameters n, p; $0 \le p \le 1$	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	$(pe^t + 1 - p)^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\frac{\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \le p \le 1$	$p(1-p)^{x-1}$ $x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial with parameters r, p; $0 \le p \le 1$	$\binom{n-1}{r-1} p^r (1-p)^{n-r}$ $n = r, r+1, \ldots$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

TABLE 7.2: CONTINUOUS PROBABILITY DISTRIBUTION

	Probability mass function, $f(x)$	Moment generating function, $M(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters $(s, \lambda), \lambda > 0$	$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)} & x \ge 0 \\ 0 & x < 0 \end{cases}$	$\left(\frac{\lambda}{\lambda - t}\right)^s$	$\frac{s}{\lambda}$	$\frac{s}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2

7.8- Additional properties of normal random variables

7.8.1- The Multivariate Normal Distribution

Reminder

Let $Z_1, \ldots, Z_n$ be a set of n independent unit normal random variables. If, for some constants $a_{ij}, 1 \le i \le m, 1 \le j \le n$, and $\mu_i, 1 \le i \le m$,

$$
\begin{aligned}
X_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n + \mu_1 \\
X_2 &= a_{21}Z_1 + \cdots + a_{2n}Z_n + \mu_2 \\
&\vdots \\
X_i &= a_{i1}Z_1 + \cdots + a_{in}Z_n + \mu_i \\
&\vdots \\
X_m &= a_{m1}Z_1 + \cdots + a_{mn}Z_n + \mu_m
\end{aligned}
$$

then the random variables $X_1, \ldots, X_m$ are said to have a multivariate normal distribution.

From the fact that the sum of independent normal random variables is itself a

7.8.2- The Joint Distribution of the Sample Mean & Sample Variance

7.9 - General Definition of Expectation

Up to this point, we have defined expectations only for discrete and continuous random variables.

However, there are also exist random variables that are neither discrete nor continuous but they may posses an expectation

Example:
Let X be Bernoulli random variable with parameters $p = \frac{1}{2}$, and let Y be a uniformly distributed random variable over the interval $[0,1]$. Furthermore, suppose that X and Y

are independent, and define the new random variable W by

$$W = \begin{cases} X & \text{if } X = 1 \\ Y & \text{if } X \neq 1 \end{cases}$$

Clearly, W is neither a discrete (since its set of possible values $[0,1]$ is uncountable), nor continuous since $P(W=1) = \frac{1}{2}$ random variable.

* In order to define the expectation of an arbitrary random variable, we require the notion of a

Stieltjes integral.

Before defining the integral, let us recall that, for any function g, $\int_a^b g(x)\,dx$ is defined by:

$$\int_a^b g(x)\,dx = \lim \sum_{i=1}^{n} g(x_i)(x_i - x_{i-1})$$

where the limit taken over all

$a = x_0 < x_1 < x_2 \cdots < x_n = b$

as $n \to \infty$ where

$\max(x_i - x_{i-1}) \to 0$

$i = 1, \ldots, n$

For any distribution function F, we define the Stieltjes integral of the non-negative function g over the interval $[a, b]$, by

$$\int_a^b g(x)\,dF(x) = \lim \sum_{i=1}^{n} g(x_i)[F(x_i) - F(x_{i-1})]$$

where, as before, the limit is taken over all $a = x_0 < x_1 < \cdots < x_n = b$ as $n \to \infty$ and where $\max_{i=1,\ldots,n} (x_i - x_{i-1}) \to 0$. Further, we define the Stieltjes integral over the whole real line by

$$\int_{-\infty}^{\infty} g(x)\,dF(x) = \lim_{\substack{a \to -\infty \\ b \to +\infty}} \int_a^b g(x)\,dF(x)$$

Finally, if g is not a nonnegative function, we define g^+ and g^- by

$$g^+(x) = \begin{cases} g(x) & \text{if } g(x) \geq 0 \\ 0 & \text{if } g(x) < 0 \end{cases}$$

$$g^-(x) = \begin{cases} 0 & \text{if } g(x) \geq 0 \\ -g(x) & \text{if } g(x) < 0 \end{cases}$$

Because $g(x) = g^+(x) - g^-(x)$ and g^+ and g^- are both nonnegative functions, it is natural to define

$$\int_{-\infty}^{\infty} g(x)\,dF(x) = \int_{-\infty}^{\infty} g^+(x)\,dF(x) - \int_{-\infty}^{\infty} g^-(x)\,dF(x)$$

and we say that $\int_{-\infty}^{\infty} g(x)\,dF(x)$ exists as long as $\int_{-\infty}^{\infty} g^+(x)\,dF(x)$ and $\int_{-\infty}^{\infty} g^-(x)\,dF(x)$ are not both equal to $+\infty$.

If X is an arbitrary random variable having cumulative distribution F, we define the expected value of X by

$$E[X] = \int_{-\infty}^{\infty} x\,dF(x) \tag{9.1}$$

It can be shown that if X is a discrete random variable with mass function $p(x)$, then

$$\int_{-\infty}^{\infty} x dF(x) = \sum_{x:p(x)>0} xp(x)$$

whereas if X is a continuous random variable with density function $f(x)$, then

$$\int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{\infty} xf(x)\,dx$$

The reader should note that Equation (9.1) yields an intuitive definition of $E[X]$; consider the approximating sum

$$\sum_{i=1}^{n} x_i[F(x_i) - F(x_{i-1})]$$

of $E[X]$. Because $F(x_i) - F(x_{i-1})$ is just the probability that X will be in the interval $(x_{i-1}, x_i]$, the approximating sum multiplies the approximate value of X when it is in the interval $(x_{i-1}, x_i]$ by the probability that it will be in that interval and then sums over all the intervals. Clearly, as these intervals get smaller and smaller in length, we obtain the "expected value" of X.

Stieltjes integrals are mainly of theoretical interest because they yield a compact way of defining and dealing with the properties of expectation. For instance, the use of Stieltjes integrals avoids the necessity of having to give separate statements and proofs of theorems for the continuous and the discrete cases. However, their properties are very much the same as those of ordinary integrals, and all of the proofs presented in this chapter can easily be translated into proofs in the general case.

SUMMARY

SUMMARY

If X and Y have a joint probability mass function $p(x, y)$, then

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$$

whereas if they have a joint density function $f(x, y)$, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\, dx\, dy$$

A consequence of the preceding equations is that

$$E[X + Y] = E[X] + E[Y]$$

which generalizes to

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

The *covariance* between random variables X and Y is given by

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

A useful identity is

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

When $n = m$ and $Y_i = X_i, i = 1, \ldots, n$, the preceding formula gives

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j} \sum \text{Cov}(X_i, Y_j)$$

The correlation between X and Y, denoted by $\rho(X, Y)$, is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

If X and Y are jointly discrete random variables, then the conditional expected value of X, given that $Y = y$, is defined by

$$E[X|Y = y] = \sum_x xP\{X = x|Y = y\}$$

If X and Y are jointly continuous random variables, then

$$E[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)$$

where

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

is the conditional probability density of X given that $Y = y$. Conditional expectations, which are similar to ordinary expectations except that all probabilities are now computed conditional on the event that $Y = y$, satisfy all the properties of ordinary expectations.

Let $E[X|Y]$ denote that function of Y whose value at $Y = y$ is $E[X|Y = y]$. A very useful identity is

$$E[X] = E[E[X|Y]]$$

In the case of discrete random variables, this equation reduces to the identity

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\}$$

and, in the continuous case, to

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)$$

The preceding equations can often be applied to obtain $E[X]$ by first "conditioning" on the value of some other random variable Y. In addition, since, for any event A, $P(A) = E[I_A]$, where I_A is 1 if A occurs and is 0 otherwise, we can use the same equations to compute probabilities.

The conditional variance of X, given that $Y = y$, is defined by

$$\text{Var}(X|Y = y) = E[(X - E[X|Y = y])^2|Y = y]$$

Let $\text{Var}(X|Y)$ be that function of Y whose value at $Y = y$ is $\text{Var}(X|Y = y)$. The following is known as the *conditional variance formula*:

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

Suppose that the random variable X is to be observed and, on the basis of its value, one must then predict the value of the random variable Y. In such a situation, it turns out that, among all predictors, $E[Y|X]$ has the smallest expectation of the square of the difference between it and Y.

The *moment generating function* of the random variable X is defined by

$$M(t) = E[e^{tX}]$$

The moments of X can be obtained by successively differentiating $M(t)$ and then evaluating the resulting quantity at $t = 0$. Specifically, we have

$$E[X^n] = \frac{d^n}{dt^n}M(t)\bigg|_{t=0} \qquad n = 1, 2, \ldots$$

Two useful results concerning moment generating functions are, first, that the moment generating function uniquely determines the distribution function of the random variable and, second, that the moment generating function of the sum of independent random variables is equal to the product of their moment generating functions. These results lead to simple proofs that the sum of independent normal (Poisson, gamma) random variables remains a normal (Poisson, gamma) random variable.

If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables, then their *sample mean*

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

and their *sample variance*

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n - 1}$$

are independent. The sample mean $\overline{X}$ is a normal random variable with mean μ and variance σ^2/n; the random variable $(n - 1)S^2/\sigma^2$ is a chi-squared random variable with $n - 1$ degrees of freedom.